Exercise 1: Risk Minimizers for 0-1-Loss

Consider the classification learning setting, i.e., $\mathcal{Y} = \{1, \ldots, g\}$, and the hypothesis space is $\mathcal{H} = \{h : \mathcal{X} \to \mathcal{Y}\}$. The loss function of interest is the 0-1-loss:

$$L(y, h(\mathbf{x})) = \mathbb{1}_{\{y \neq h(\mathbf{x})\}} = \begin{cases} 1, & \text{if } y \neq h(\mathbf{x}), \\ 0, & \text{if } y = h(\mathbf{x}). \end{cases}$$

(a) Consider the hypothesis space of constant models $\mathcal{H} = \{h : \mathcal{X} \to \mathcal{Y} \,|\, h(\mathbf{x}) = \boldsymbol{\theta} \in \mathcal{Y} \ \forall \mathbf{x} \in \mathcal{X}\}$. where $\mathcal{X}$ is the feature space. Show that

$$\hat{h}(\mathbf{x}) = \text{mode}\left\{y^{(i)}\right\}$$

is the empirical risk minimizer for the 0-1-loss in this case.

(b) What is the optimal constant model in terms of the (theoretical) risk for the 0-1-loss and what is its risk?

(c) Derive the approximation error if the hypothesis space $\mathcal{H}$ consists of the constant models.

(d) Assume now $g = 2$ (binary classification) and consider now the hypothesis space of probabilistic classifiers $\mathcal{H} = \{\pi : \mathcal{X} \to [0, 1]\}$, that is, $\pi(\mathbf{x})$ (or $1 - \pi(\mathbf{x})$) is an estimate of the posterior distribution $p_{y|x}(1 \mid \mathbf{x})$ (or $p_{y|x}(0 \mid \mathbf{x})$). Further, consider the probabilistic 0-1-loss

$$L(y, \pi(\mathbf{x})) = \begin{cases} 1, & \text{if } (\pi(\mathbf{x}) \geq 1/2 \text{ and } y = 0) \text{ or } (\pi(\mathbf{x}) < 1/2 \text{ and } y = 1), \\ 0, & \text{else.} \end{cases}$$

Is the minimum of $\mathbb{E}_{xy}[L(y, \pi(\mathbf{x}))]$ unique over $\pi \in \mathcal{H}$[1]? Is the posterior distribution $p_{y|x}$ a resp. *the* minimizer of $\mathbb{E}_{xy}[L(y, \pi(\mathbf{x}))]$? Discuss the corresponding (dis-)advantages of your findings.

Hint: First note that we can write $L(y, \pi(\mathbf{x})) = \mathbb{1}_{\{\pi(\mathbf{x}) \geq 1/2\}} \mathbb{1}_{\{y=0\}} + \mathbb{1}_{\{\pi(\mathbf{x}) < 1/2\}} \mathbb{1}_{\{y=1\}}$ and then consider the "unraveling trick": $\mathbb{E}_{xy}[L(y, \pi(\mathbf{x}))] = \mathbb{E}_x\left[\mathbb{E}_{y|x}[L(y, \pi(\mathbf{x})) \mid \mathbf{x} = \mathbf{x}]\right]$.

[1] If it is unique, then the loss is a strictly proper scoring rule.